JOURNAL OF PLANKTON RESEARCH | VOLUME 31 | NUMBER 1 | PAGES 19–29 | 2009

Optimizing the number of classes in automated zooplankton classification

JOSE A. FERNANDES[1,2]*, XABIER IRIGOIEN[1], GUILLERMO BOYRA[1], JOSE A. LOZANO[2] AND IÑAKI INZA[2]

[1]AZTI-TECNALIA, MARINE RESEARCH DIVISION, HERRERA KAIA PORTUALDEA Z/G, E-20110 PASAIA, SPAIN AND [2]DEPARTMENT OF COMPUTER SCIENCE AND AI, UNIVERSITY OF THE BASQUE COUNTRY, INTELLIGENT SYSTEMS GROUP (ISG), PASEO MANUEL DE LARDIZABAL, 1, E-20018 DONOSTIA–SAN SEBASTIÁN, SPAIN

*CORRESPONDING AUTHOR: jfernandes@pas.azti.es

Received March 11, 2008; accepted in principle September 12, 2008; accepted for publication September 21, 2008; published online 22 October, 2008

Corresponding editor: Roger Harris

Zooplankton biomass and abundance estimation, based on surveys or time-series, is carried out routinely. Automated or semi-automated image analysis processes, combined with machine-learning techniques for the identification of plankton, have been proposed to assist in sample analysis. A difficulty in automated plankton recognition and classification systems is the selection of the number of classes. This selection can be formulated as a balance between the number of classes identified (zooplankton taxa) and performance (accuracy; correctly classified individuals). Here, a method is proposed to evaluate the impact of the number of selected classes, in terms of classification performance. On the basis of a data set of classified zooplankton images, a machine-learning method suggests groupings that improve the performance of the automated classification. The end-user can accept or reject these mergers, depending on their ecological value and the objectives of the research. This method permits both objectives to be equally balanced: (i) maximization of the number of classes and (ii) performance, guided by the end-user.

INTRODUCTION

The study of zooplankton abundance and biomass distribution is important, in order to understand marine ecosystems. Although a routine task in many laboratories, it still presents a practical challenge to marine scientists. Furthermore, the temporal and spatial sampling scales required to understand distribution (Mackas, 1984; Steele, 1989) are incompatible with laborious sample analysis using a microscope. In order to bring zooplankton research to the same level of spatial and temporal resolution as that of phytoplankton (Chl *a*), a wide range of image analysis and automatic recognition methods have been proposed (Benfield *et al.*, 2007). Image analysis, combined with automatic classification, offers a number of advantages in terms of speed of analysis, replication and error estimation (Culverhouse *et al.*, 2003; Benfield *et al.*, 2007). Furthermore, it can achieve accuracies (correctly classified rates) comparable with those achieved by humans, for a number of classes (e.g. taxa, artifacts, size-based groups), which are of ecological significance (Culverhouse *et al.*, 2003; Grosjean *et al.*, 2004).

An exhaustive list of imaging devices and sample digitalizing approaches can be found in Culverhouse *et al.* (2006) and Benfield *et al.* (2007). Once the samples have been digitalized, image analysis is used to separate automatically the different images of individuals present in each sample. In this step, several measurements or features are extracted to represent each of those individuals (Table I). A comprehensive conceptual diagram of the image analysis and posterior automated classification process is provided by Fig. 7a of Benfield *et al.* (2007). In the automated classification step, the user has to label (classify) manually a representative fraction of the individuals. These labeled individuals are used to build a classifier (a computer algorithm or method). This classifier is then automatically used on the remaining unlabeled individuals (thousands to millions of them). Such machine-learning concepts are described in Alpaydin (2004).

doi:10.1093/plankt/fbn098, available online at www.plankt.oxfordjournals.org

Table I: Individual features: Morphological and image measurements extracted by ZooImage, using the image analysis software, ImageJ

Feature	Description
ZooImage (ImageJ) features	
ECD	Equivalent circular diameter
Area	Surface area
Mean	Mean of the gray scale of the pixels
Skew	The third-order moment, about the mean of the gray scale
Kurt	The fourth-order moment, about the mean of the gray scale
StdDev	Standard deviation of the gray scale of the pixels
Mode	Mode of the gray scale of the pixels
Median	Median of the gray scale of the pixels
Min	Minimum of the gray scale of the pixels
Max	Maximum of the gray scale of the pixels
IntDen	Sum of the gray values of the pixels
XM	Coordinate horizontal of the gray scale center of the pixels
YM	Coordinate vertical of the gray scale center of the pixels
Perim.	Perimeter
Width	Width of the rectangle, containing the individual
Height	Height of the rectangle, containing the individual
Major	Longest axis of the ellipsis, containing the individual
Minor	Smallest axis of the ellipsis, containing the individual
Circ.	Circularity
Feret	Diameter of longest distance between the two points of the individual
Environmental features	
Temperature	Surface temperature
Salinity	Salinity of the sample
Depth	Depth of the sample
Latitude	Latitude of the sample
Longitude	Longitude of the sample

Environmental features collected during the survey have been added after image analysis to DataSet2.

For both humans and machine-learning techniques, there is a "trade-off" between the number of classes to be identified and classification performance (Culverhouse *et al.*, 2003). On the one hand, an increase in the number of classes usually leads to a decrease in performance. It is harder to distinguish between classes and easier to make labeling mistakes; on the other hand, the aggregation of groups can bring about an increase in performance. One of the first challenges faced by the "end-user" who is undertaking the labeling is whether to create a new group for an individual or to use a previously defined one. A large number of classes is desirable for several reasons: (i) the more ecologically meaningful classes that can be separated, the more the understanding of the system is improved; (ii) particle biomass contribution should be calculated using different conversion factors depending on the group they belong to (Strathmann, 1967; Wiebe *et al.*, 1975; Alcaraz *et al.*, 2003); (iii) in imaging systems, there are particles (such as artifacts or bubbles), whose contribution to biomass should be assumed to be zero. However, should there be too many classes, the classification error can be excessively high and the whole process undermined. Therefore, during labeling, the end-user seeks a balance between the number of classes and accuracy. Such a search is undertaken manually, in a lengthy and uncertain trial and error process.

Once the number of classes has been set, a classifier is evaluated to identify the expected performance. At the same time, the confusion matrix (CM) is generated (Luo *et al.*, 2004; Hu and Davis, 2006). A CM is a graphical representation that compares the user classification with the classifier classification, showing how the error has been distributed (Table II). An end-user can examine the incorrectly classified individuals in the CM, to assess the conflicting classes and undertake re-labeling. However, this leads to a manual "trial and error" loop of building classifiers and CMs, i.e. aggregating these conflicting classes, or labeling more individuals, without any certainty of achieving an improved performance. As such, it is not easy to decide which classes to group and to envisage the impact of these actions on the discrimination between classes. From a human perspective, it may be difficult to establish which classes can be differentiated and which could cause confusion (for machine recognition). Humans and machines do not necessarily utilize the same classification factors.

To the best of our knowledge, this class selection problem has not yet been solved; it is left to the end-user, in the published literature. Actual machine-learning techniques are more concerned with merging images that are similar in shape (clustering techniques), than their meaning and performance (Donamukkala *et al.*, 2005). In plankton studies, morphologically similar classes can be different from an ecological perspective and should not be aggregated. Therefore, the best class grouping for an automatic classifier need not necessarily be appropriate for an end-user to extract meaningful information. A balance, taking both points of view into consideration, is required.

We propose a method that combines human knowledge with machine-learning techniques in order to allow the end-user to determine if the performed labeling, in terms of number of classes, can be improved on or not. The aim is to maximize both the performance of the classifier and the number of classes with robust and meaningful information for the end-user. A machine-learning method provides the statistics of performance and the number of classes, whereas the end-user provides the ecologically meaningful information and the initial number of classes.

Table II: CM of the classifier before mergers evaluation for DataSet1

User ↓ | Classifier →	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
Oncaeidae (A)	**29**	0	3	4	0	8	0	0	0	0	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
Round Egg (B)	0	**38**	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	1	1	0	2	0	0	1	1
Calanoida Lateral (C)	6	0	**12**	1	0	5	0	0	0	0	8	1	0	1	1	2	3	0	0	0	0	0	0	0	0	3	2	2	2	0	0	0	0	0	1	0	0
Corycaeidae (D)	1	0	3	**18**	0	12	1	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	8	0	4	0	0	0	0	0	0	0	0	0
Miraciidae (E)	0	0	0	0	**38**	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Poicilo Lateral (F)	9	0	0	13	0	**18**	0	0	0	0	2	0	0	1	0	0	1	0	0	0	0	0	0	0	0	3	0	1	2	0	0	0	0	0	0	0	0
Eucalanidae (G)	0	0	0	1	0	0	**40**	0	0	3	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
Scratch (H)	0	0	0	0	0	0	0	**47**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Debris (I)	0	3	0	0	0	0	0	0	**24**	0	0	0	0	0	0	0	0	0	5	0	0	0	0	1	0	6	2	0	0	0	1	3	4	0	0	0	1
Calanoida Dorsal III (J)	0	0	0	0	1	0	7	0	0	**34**	0	5	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Calanoida Dorsal I (K)	2	0	5	1	0	1	3	0	0	0	**24**	5	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	5	0	0
Calanoida Dorsal II (L)	3	0	1	0	0	0	1	0	0	6	3	**22**	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	1	9	0	0
Fiber (M)	0	0	0	0	1	0	1	1	0	0	0	0	**40**	0	0	0	0	0	0	0	4	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0
Decapoda Miscellaneous (N)	0	0	0	0	0	0	0	0	1	0	2	0	0	**11**	0	0	1	0	0	0	0	0	0	0	0	0	7	4	0	0	0	0	0	0	0	0
Appendicularia (O)	0	0	1	0	0	0	0	0	0	0	0	2	1	0	**21**	0	0	0	2	0	5	2	0	0	0	0	0	1	0	0	2	0	0	0	0	0
Decapoda Zoea Lateral (P)	2	0	1	0	0	1	0	0	0	0	1	0	0	1	0	**27**	5	0	0	0	0	0	0	0	0	2	0	4	4	0	2	0	0	0	0	0
Decapoda Zoea Dorsal (Q)	3	0	0	0	0	1	0	0	1	0	0	0	0	0	0	6	**30**	0	0	0	0	0	0	0	0	1	0	5	1	0	1	0	0	0	1	0	0
Bubble (R)	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	**16**	0	0	0	1	2	1	3	0	0	0	0	0	1	0	0	0	0	0	0
Annelida (S)	0	0	2	0	1	0	2	0	3	1	2	1	0	1	0	3	0	0	**24**	0	0	0	0	0	0	1	0	1	3	1	0	0	1	0	0	3	0
Sapphirinidae (T)	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	1	0
Oithonidae (U)	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2	0	0	0	0	0	**40**	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Cirripeda (V)	0	0	2	0	0	0	0	0	0	0	1	0	1	0	2	0	1	0	0	0	3	**5**	0	0	2	0	0	0	1	0	0	2	0	0	2	0	0
Shadow (W)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	**47**	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Diatom (X)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**43**	5	0	0	0	0	0	0	2	0	0	0	0	0
Protista (Y)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	**47**	0	0	0	0	0	0	0	0	0	0	0	0
Pisces (Z)	4	2	1	1	0	2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**37**	0	0	0	0	1	0	0	0	0	0	0
Elongated Egg (AA)	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**48**	0	0	0	0	0	0	0	0	0	0
Malacostraca Bulky (AB)	0	0	1	1	0	0	0	0	0	0	0	0	0	5	0	0	0	0	1	0	0	0	0	0	0	0	0	**38**	2	0	0	0	0	0	1	1	0
Marine Snow (AC)	1	0	4	0	0	2	0	0	2	0	3	1	0	5	1	2	0	0	2	0	0	0	0	0	0	1	0	6	**15**	0	0	0	0	1	1	3	0
Cnidaria (AD)	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	**14**	0	2	0	1	0	1	2
Cladocera (AE)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	0	0	0	1	0	0	0	0	2	0	0	0	0	**41**	0	0	0	0	0	0
Other Phytoplankton (AF)	0	0	0	0	0	0	2	0	0	0	0	0	6	0	3	0	0	0	0	0	1	0	1	1	0	1	0	0	2	1	0	**29**	0	0	0	0	3
Gastropoda (AG)	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	**44**	0	0	0	0
Malacostraca Larvae (AH)	4	0	6	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2	0	0	0	3	0	0	2	0	0	0	0	**1**	0	0	0
Temoridae (AI)	2	0	1	0	0	0	3	0	0	0	8	17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**7**	0	0
Elongated Malacostraca (AJ)	0	0	2	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	**44**	0
Chaetognatha (AK)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	**48**

The main diagonal in the center represents the correctly classified individuals. The rest of the cells are the misclassified individuals. Columns show the classifier classification and rows show the user labeled class present in the data set. The selected mergers by the end-user are displayed in gray.

METHOD AND RESULTS

Procedure

The methodology proposed consists of three steps, as outlined below.

(1) The end-user distributes the extracted images of individuals into all the groups, which can be visually identified (i.e. labeling). A classifier is trained with this data set and the corresponding estimated performance is used as a starting point, which will be improved on subsequently. Any methodology and tools for data acquisition preferred by the expert can be used.

(2) All possible mergers of two classes, as a single class ("a merger"), are evaluated. For each pair of classes, a new data set is constructed, in which the two classes are merged into a unique class, whereas the remainder are left unchanged. A classifier is constructed from this new data set and its performance is evaluated. The possible mergers are ranked, based on their estimated performance (e.g. accuracy). Optionally, the CM can be used to reduce the number of mergers to be evaluated using the classes with more misclassified individuals' counts above a certain threshold (e.g. mean of non-zero misclassified; Table II). This option significantly reduces computation time. Step 2 is automatically performed by a computer program (Table III), which outputs a ranking (Table IV) with all possible class configurations (mergers of two classes) and their associated statistics (see below). The Java program uses Weka API machine-learning algorithms (Witten and Frank, 2005). In order to ensure reproducibility (Buckheit and Donoho, 1995), a Java implementation of the method is available from the ISG group webpage (www.sc.ehu.es/ccwbayes/members/jafernandes/).

(3) The end-user evaluates the ranking and decides which specific mergers to accept considering not only the performance that can be achieved, but also the ecological value and the objective of the research. A new classifier, with end-user selected mergers, is trained and evaluated. This new classifier can be compared with those established in the first step (see above) and in previous iterations. The end-user can perform steps 2 and 3, repeatedly.

The method proposed relates to optimizing the number of classes (class selection) and the classification

Table III: Pseudocode of the method used to describe the method

Method pseudocode
1: While User does not end mergers evaluation
2: Build classifier before mergers
3: Evaluate classifier
4: Calculate metrics (accuracy, . . .)
5: Save classifier metrics in mergers ranking
6: For all i∈{CLASS 1, CLASS 2, . . ., CLASS *n*-1}
7: For all j∈{CLASS i+1, . . ., CLASS n}
8: If ((CM) and (CLASS i and CLASS in CM list)) or (not CM) them
9: Reset data set to original without mergers
10: Merge CLASS i and CLASS j in data set
11: Build classifier with merged data set
12: Evaluate classifier
13: Calculate metrics (accuracy, . . .)
14: Save classifier metrics in mergers ranking
15: End If
16: End For
17: End For
18: Perform user selected mergers
19: End While

Pseudocode is not language-programming dependent; and it omits programming details that are not relevant to specify the method. CM represents if the use of confusion matrix has been selected or not.

Table IV: Ranking of mergers with highest accuracies for DataSet1, before any merger

Top 10 mergers iteration 1	User decision
66.5%, PRE: 5.1%, Oncaeidae with Calanoida Lateral	X: One is Poecilostomatoida, the other Calanoida
66.3%, PRE: 4.5%, Corycaeidae with Poicilo Lateral	√: Both are Poecilostomatoida
66.2%, PRE: 4.2%, Scratch with Temoridae	X: Scratch is an artifact and Temoridae is not
66.0%, PRE: 3.7%, Oncaeidae with Poicilo Lateral	√: Both are Poecilostomatoida
65.9%, PRE: 3.4%, Eucalanidae with Calanoida Dorsal III	√: Both are Calanoida
65.8%, PRE: 3.1%, Decapoda Miscellaneous with Decapoda Zoea Lateral	√: Both are Decapoda
65.8%, PRE: 3.1%, Decapoda Miscellaneous with Malacostraca Larvae	X: One is Decapoda and the other Malacostracea
65.7%, PRE: 2.8%, Decapoda Zoea Lateral with Pisces	X: One is Decapoda the other Pisces
65.7%, PRE: 2.8%, Decapoda Zoea Dorsal with Gastropoda	X: One is Decapoda the other is Gastropoda
65.6%, PRE: 2.5%, Decapoda Miscellaneous with Malacostraca Bulky	X: One is Decapoda and the other Malacostracea

In each row, the accuracy, the PRE, the classes to be merged with the user-decision are given.

performance. Therefore, it can be applied to data from any source and classified with different methods as long as they are classified into different classes that can be grouped without losing all the information (e.g. grouping different taxonomic levels). The method could be run "manually" but the expert would be confronted with hundreds of mergers to explore without previous knowledge of the potential accuracy gain. Any merger does not lead to an accuracy gain; in fact, there is a high rate of mergers that decrease performance (Table V). Automation and ranking of the results leave only a limited number of mergers, with higher accuracy, for the end-user to analyze; as opposed to the end-user manual "trial and error" exploration without previous knowledge of the potential performance gain.

The method is independent of any specific machine-learning paradigm for classification or evaluation and specific performance metric. The end-user can select different classification paradigms and performance metrics taking into account the specific requirements of the study being undertaken (e.g. taxonomic groups, compared with ecological impact). In our examples, a Tree Augmented Naive Bayes classifier (TAN) was used for classification (Friedman *et al.*, 1997); this has a good performance record, laying close to Random Forest, proved to be a good classification algorithm for zooplankton (Grosjean *et al.*, 2004). Indeed, the TAN is faster to be trained than Random Forest. The TAN model considers probabilistic dependencies between variables, in the form of a tree structure. The tree structure representation permits excellent computing performance, providing an intuitive and transparent representation that can be useful for the end-user to extract domain knowledge. The use of TAN is employed for a faster mergers evaluation. The final model can utilize any classification paradigm (e.g. Random Forest).

Statistics to evaluate the goodness of the new classifiers

In order to establish the expected error, the classifier performance has to be assessed; this is accomplished by dividing the user-labeled data set into two parts: training and evaluation. Depending on the selected evaluation technique, this demarcation can be undertaken once or several times, with different data sampling techniques. A popular evaluation technique is "*k*-fold cross-validation" (Stone, 1974; Geisser, 1975; Schaffer, 1993). Using this technique, the data are randomly divided into *k* parts (folds). The classifier is trained and evaluated *k* times, each time using a different fold for evaluation and the rest for training. The classifier performance is the mean of the results in the *k*-test folds. The *k* results can be used to test if differences in performance between the two different classifiers are statistically significant or due to randomness (Bouckaert and Frank, 2004). Although 10-fold and 20-fold has been proved to be the best option, since the cross-

Table V: For each iteration, several statistics are presented after performing the end-user selected mergers

Merger evaluation		DataSet1	DataSet2	DataSet3
Before	Accuracy (%)	64.7	85.7	82
After first iteration	Accuracy (%)	68.3	87.3	82.1
	P-value original	0.585	0.078	0.976
	PRE original (%)	10.2	4.7	0.6
	#Mergers selected	4	5	4
	#Mergers evaluated	666	276	435
	Mergers↓ (%)	78.3	21.4	91
	CPU-time	3:01:39	0:32:34	1:30:47
	CPU-time CM	0:17:37	0:16:07	0:17:31
	#Mergers evaluated CM	58	29	33
After second iteration	Accuracy (%)	70.9	88.8	–
	P-value previous	0.542	0.7	–
	P-value original	0.395	0.006	–
	PRE previous (%)	8.2	4.6	–
	PRE original (%)	17.6	9	–
	#Mergers selected	4	1	–
	#Mergers evaluated	528	190	–
	Mergers↓ (%)	74.7	63.7	–
	CPU-time	1:57:40	0:17:45	–
After third iteration	Accuracy (%)	73	–	–
	P-value previous	0.514	–	–
	P-value original	0.179	–	–
	PRE previous (%)	7.2	–	–
	PRE original (%)	23.5	–	–
	#Mergers selected	2	–	–
	Mergers↓ (%)	69	–	–
	CPU-time	1:41:29	–	–
After fourth iteration	Accuracy (%)	73.9	–	–
	P-value previous	0.426	–	–
	P-value original	0.699	–	–
	PRE previous (%)	3.3	–	–
	PRE original (%)	26.1	–	–
	#Mergers selected	1	–	–
	Mergers↓ (%)	16.9	–	–
	CPU-time	1:01:32	–	–
After fifth iteration	Accuracy (%)	74	–	–
	P-value previous	0.679	–	–
	P-value original	0.398	–	–
	PRE previous (%)	0.4	–	–
	PRE original (%)	26.3	–	–
	#Mergers selected	1	–	–
	Mergers↓ (%)	20.4	–	–
	CPU-time	0:40:37	–	–

"Before" represents accuracies before performing any merge. The number of evaluated mergers is represented by "#Mergers". Accuracy is the overall accuracy after performing selected mergers. "*P*-value original" is the result of performing a paired *t*-test, with the original data set before any merger, whereas "*P*-value previous" is the test but with the resulting data set of the previous iteration. The same with PRE that is provided both in relation with the previous iteration data set and in relation with the original. "Mergers↓" is the rate of mergers that instead of improving accuracy reduces it. "CPU-time" is the computer time to evaluate the mergers. "CM" corresponds to statistics when using the confusion matrix to reduce the mergers to be evaluated.

validation has to be performed hundreds of times, 5-fold has been considered sufficient to suggest the mergers (Kohavi, 1995).

In order to assess the classifier performance, several additional measures are used; percent reduction in error (PRE), true positive per class (TP), false positive per class (FP) and whether the classifiers' accuracy is significantly different (corrected paired *t*-test; Nadeau and Bengio, 2003). However, this test is conservative and can result in higher *P*-values than other less strict tests (e.g. paired *t*-test). Accuracy, overall correctly classified, is used as the main metric because it is a simple way of assessing performance (Pazzani, 1996; Kohavi and John, 1997). However, the end-user can define other metrics depending on the study objectives. Finally, the relevance of a performance gain can be hard to understand. For example, a 2% accuracy gain on an already high accuracy (e.g. 90%) is not the same as with a low accuracy (e.g. 50%). This can be measured using the PRE (Hagle and Glen, 1992). PRE = (100 × (EB − EA)/EB), where EB is error before mergers and EA error after mergers. TP is the proportion of individuals that have

been correctly classified as belonging to a class. FP is the proportion of individuals that not being of a certain class are incorrectly classified as being part of it.

Application examples

In order to illustrate the method, we have applied it to three different data sets: a public data set (DataSet1) available at the ZooImage webpage (www.sciviews.org/zooimage) has been selected in order to permit reproducibility. In addition, we have used two data sets obtained at different imaging resolution (Table V). DataSet2 has been established with zooplankton samples scanned at 600 dpi; DataSet3 with 2400 dpi images. Both data sets have been built from samples obtained in the Bay of Biscay preserved in 4% borax buffered formalin, then stained with eosin. Eosin staining avoids the imaging of inorganic debris in the image analysis step through image filters. Both data sets were analyzed using ZooImage. The variables considered were those routinely extracted by ZooImage, together with a limited number of environmental variables for DataSet2 (Table I). In DatSet1, 1639 individuals were classified into 37 classes. For the DataSet2, 17803 individuals were classified into 24 classes. For DataSet3, 6724 were classified into 30 classes (Table VI, Fig. 1). The data sets are used for illustration purposes; the method can be applied to data sets obtained with any other methodology.

The evaluation of the new classifiers with class mergers is shown in Tables V and VII. In DataSet1, 64.7% accuracy was obtained with the 37 classes. Out of 666 possible two-class mergers considered, 145 (21.7%) showed an improvement in accuracy. The list of

Table VI: Number of individuals per class in the different data sets, before any merger

DataSet1		DataSet2		DataSet3	
Number of individuals	Classes	Number of individuals	Classes	Number of individuals	Classes
27	Bubble	467	Artifact	110	Artifact
50	Scratch	482	Small Marine Snow	97	Small Marine Snow
50	Shadow	1136	Marine Snow	97	Medium Marine Snow
50	Debris	2228	Small Copepoda	49	Large Marine Snow
50	Diatom	2063	Medium Copepoda	198	Small Copepoda
50	Fiber	2361	Large Copepoda	207	Copepoda multiple
50	Marine Snow	871	Multiple Copepoda	2288	Calanoida
50	Other Phytoplankton	1838	Euphausiacea	1189	Cyclopoida Oncaea
50	Calanoida Dorsal I	208	Decapoda Larvae	110	Cyclopoida Corycaeus
49	Calanoida Dorsal II	122	Decapoda Larvae II	548	Cyclopoida Oithona_sp
50	Calanoida Dorsal III	279	Polychaeta	86	Cyclopoida Oithona_nana
50	Calanoida Lateral	12	Polychaeta Larvae I	168	Harpaticoida Microsetella
50	Eucalanidae	31	Amphipoda	208	Harpaticoida Euterpina
39	Temoridae	209	Appendicularia	174	Appendicularia
50	Oithonidae	1123	Chaetognatha	115	Chaetognatha
39	Miraciidae	107	Doliolida	12	Euphausiacea
50	Corycaeidae	202	Siphonophorae	32	Decapoda Larvae
50	Oncaeidae	57	Hydroidomedusae	244	Cladocera
50	Poicilo Lateral	160	Stained Jelly (rests)	28	Nematoda
8	Sapphirinidae	17	Cephelopoda Larvae	250	Doliolid
50	Annelida	48	Pisces	20	Siphonophora
22	Cirripeda	200	Pisces Larvae	84	Hydroidomedusae
50	Cladocera	3043	Zooplankton small	142	Bivalvia Larvae
26	Decapoda Miscellaneous	539	Round zooplankton	18	Gastropoda
50	Decapoda Zoea Dorsal			58	Pteropoda
50	Decapoda Zoea Lateral			32	Polychaeta
50	Malacostraca Bulky			52	Copepod Egg I
50	Elongated Malacostraca			46	Copepod Egg II
21	Malacostraca Larvae			16	Fish Egg
22	Cnidaria			16	Diatom
37	Appendicularia				
50	Chaetognatha				
50	Elongated Egg				
49	Round Egg				
50	Protista				
50	Gastropoda				
50	Pisces				
1639	37	17 803	24	6694	30

Stained Jelly class represents partial gelatinous individuals that can not be identified (See Fig. 1).

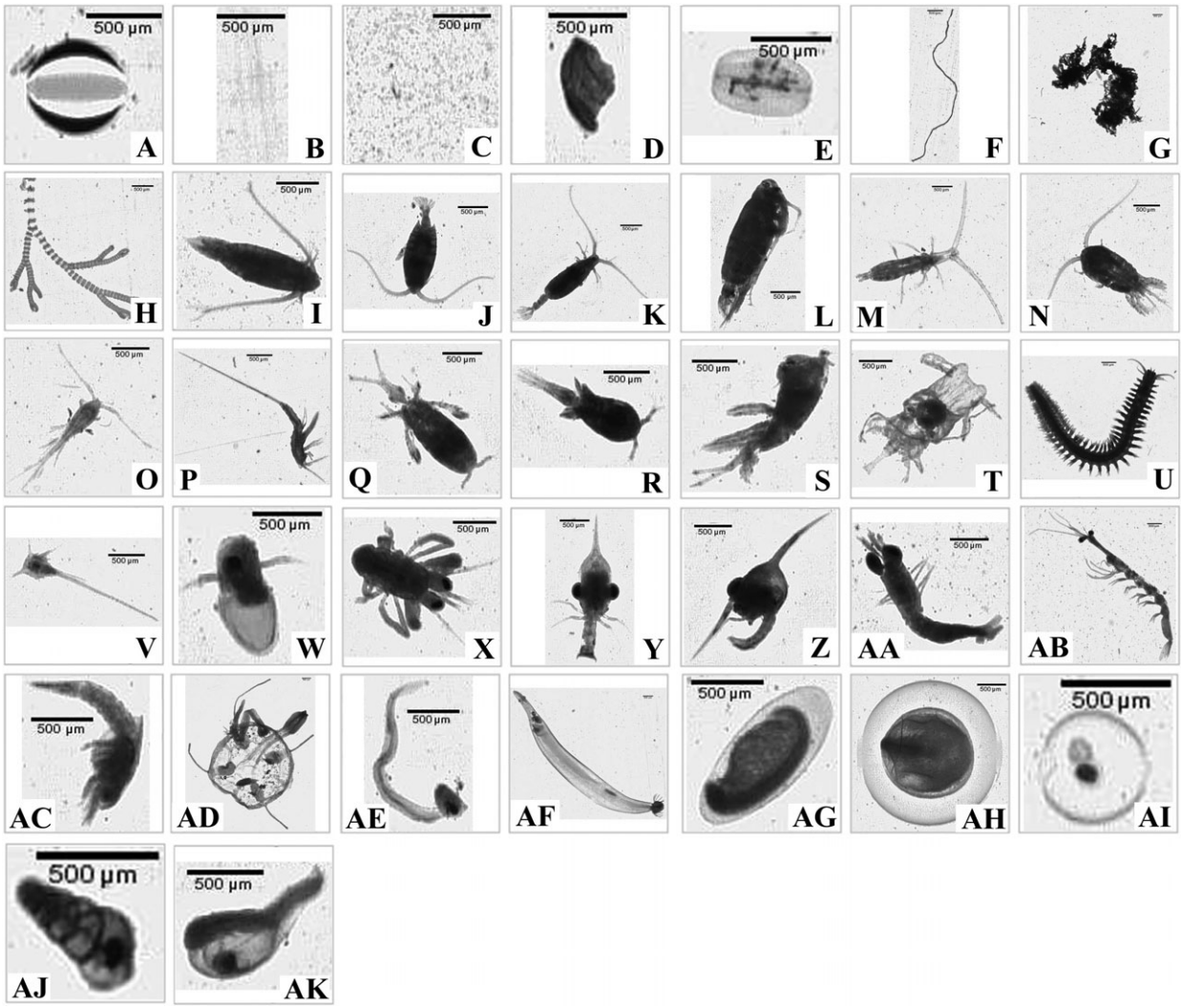

Fig. 1. Images representative of each class presented in the original DataSet1. Bubble (A), Scratch (B), Shadow (C), Debris (D), Diatom (E), Fiber (F), Marine Snow (G), Other Phytoplankton (H), Calanoida Dorsal I (I), Calanoida Dorsal II (J), Calanoida Dorsal III (K), Calanoida Lateral (L), Eucalanidae (M), Temoridae (N), Oithonidae (O), Miraciidae (P), Corycaeidae (Q), Oncaeidae (R), Poicilo Lateral (S), Sapphirinidae (T), Annelida (U), Cirripeda (V), Cladocera (W), Decapoda Miscellaneous (X), Decapoda Zoea Dorsal (Y), Decapoda Zoea Lateral (Z), Malacostraca Bulky (AA), Elongated Malacostraca (AB), Malacostraca Larvae (AC), Cnidaria (AD), Appendicularia (AE), Chaetognatha (AF), Elongated Egg (AG), Round Egg (AH), Protista (AI), Gastropoda (AJ) and Pisces (AK). See Table VI. Public available images from ZooImage web page: www.sciviews.org/zooimage.

the mergers, which resulted in a higher improvement in accuracy, was evaluated by a human end-user who accepted five mergers (Fig. 2). Several iterations were performed until there were no further mergers accepted by the user. After the third iteration, there was an 8.3% accuracy gain, a PRE of 23.5%. However, the improvement in accuracy may not be significant enough ($P < 0.20$). In DataSet2, there is a 3.1% accuracy gain after two iterations, a PRE of 9% and the classifier with mergers was significantly improved ($P < 0.05$). In DataSet3, there is a small accuracy gain of only 0.1% and the new classifier is not significantly different from the previous one ($P > 0.05$).

DISCUSSION

The proposed method consists of a semi-automated investigation of possible mergers, which can balance both objectives, i.e. the maximization of class number and the performance. The exhaustive study of all possible class combinations is computationally unfeasible, e.g. the number of possible combinations for DataSet1 (37 classes) is 3.74409×10^{43}. The total number of class mergers to be evaluated (two-classes mergers + three-classes mergers + four-classes mergers + ... + $(n-1)$-classes mergers) can be calculated by means of Stirling numbers of second kind (Abramowitz and Stegun, 1965):

Table VII: Classifier overall accuracy (correctly classified), TP rate and FP, per class in each classifier (generated after "end-user" selected mergers, in each iteration)

	Before mergers		After iteration 1		After iteration 2		After iteration 3		After iteration 4		After iteration 5	
Accuracy	0.647		0.683		0.697		0.73		0.739		0.74	
PRE	–		0.102		0.142		0.235		0.261		0.263	
Classes	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Bubble	0.593	0	0.556	0.002	0.63	0.001	0.63	0.003	0.593	0.003	0.481	0
Scratch	0.94	0.001	0.96	0.002	0.96	0.001	0.96	0.002	0.94	0.001	0.95	0.001
Shadow	0.94	0.004	0.86	0.002	0.9	0.001	0.84	0.001				
Debris	0.48	0.009	0.54	0.006	0.54	0.009	0.56	0.006	0.52	0.006	0.52	0.009
Diatom	0.86	0.003	0.86	0.003	0.88	0.004	0.86	0.003	0.88	0.004	0.88	0.004
Fiber	0.8	0.006	0.82	0.008	0.71	0.022	0.73	0.025	0.75	0.022	0.77	0.025
Other Phytoplankton	0.58	0.01	0.52	0.01								
Marine Snow	0.3	0.02	0.28	0.014	0.3	0.013	0.24	0.013	0.28	0.013	0.26	0.013
Calanoida Dorsal I	0.48	0.022	0.36	0.022	0.41	0.038	0.813	0.069	0.816	0.064	0.806	0.066
Calanoida Lateral	0.24	0.021	0.18	0.012								
Calanoida Dorsal II	0.449	0.023	0.408	0.021	0.648	0.026						
Temoridae	0.179	0.013	0.282	0.014								
Calanoida Dorsal III	0.68	0.009	0.86	0.024	0.87	0.021						
Eucalanidae	0.8	0.015										
Oithonidae	0.8	0.012	0.72	0.011	0.6	0.01	0.48	0.008	0.62	0.006	0.73	0.006
Miraciidae	0.974	0.002	0.923	0.002	0.897	0.001	0.897	0.003	0.846	0.001		
Corycaeidae	0.36	0.014	0.847	0.052	0.813	0.044	0.8	0.056	0.827	0.058	0.807	0.054
Oncaeidae	0.58	0.023										
Poicilo Lateral	0.36	0.021										
Sapphirinidae	0	0	0	0	0	0.001	0	0	0	0	0	0
Annelida	0.48	0.009	0.54	0.006	0.5	0.005	0.5	0.006	0.5	0.006	0.5	0.006
Cirripeda	0.227	0.004	0.318	0.005	0.273	0.006	0.273	0.003	0.227	0.003	0.318	0.004
Cladocera	0.82	0.004	0.86	0.006	0.82	0.004	0.84	0.004	0.84	0.004	0.84	0.004
Decapoda Miscellaneous	0.423	0.01	0.539	0.025	0.651	0.034	0.651	0.03	0.667	0.03	0.659	0.029
Decapoda Zoea Lateral	0.54	0.012										
Decapoda Zoea Dorsal	0.6	0.011	0.56	0.009								
Malacostraca Bulky	0.76	0.02	0.76	0.016	0.74	0.014	0.7	0.014	0.76	0.014	0.78	0.015
Elongated Malacostraca	0.88	0.008	0.9	0.008	0.92	0.006	0.9	0.004	0.9	0.005	0.9	0.004
Malacostraca Larvae	0.048	0.002	0.095	0.001	0	0.001	0.095	0.001	0	0.001	0.095	0.001
Cnidaria	0.636	0.003	0.591	0.005	0.636	0.004	0.591	0.006	0.545	0.005	0.591	0.006
Appendicularia	0.568	0.007	0.514	0.009	0.514	0.006	0.514	0.009	0.514	0.009	0.514	0.006
Chaetognatha	0.96	0.004	0.92	0.004	0.94	0.004	0.94	0.004	0.96	0.004	0.92	0.003
Elongated Egg	0.96	0.003	0.98	0.004	0.98	0.003	0.98	0.004	0.96	0.003	0.96	0.003
Round Egg	0.776	0.004	0.755	0.004	0.755	0.004	0.755	0.003	0.776	0.003	0.776	0.003
Gastropoda	0.88	0.004	0.84	0.004	0.86	0.004	0.88	0.003	0.86	0.004	0.88	0.004
Protista	0.94	0.007	0.94	0.005	0.94	0.006	0.94	0.005	0.92	0.004	0.96	0.005
Pisces	0.74	0.021	0.7	0.015	0.6	0.014	0.56	0.013	0.52	0.011	0.56	0.012

TP rate is the percentage of individuals classified in a class by the classifier, which belong to that class in the training set. FP is the percentage of individuals classified as belonging to a class when they are not. TP and FP experiment low variation in classes not being merged and high improvement in most of the merged classes. Figure 2 for selected mergers during each iteration.

$X = \left(\sum_{k=0}^{n} S(n,k)\right) - 2$ (excluding "not performing any merger" and "merging in a unique class"). In this expression, X is the number of possible combinations, n is the number of classes to consider for possible mergers and $S(n, k)$ is broken down as

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^n.$$

This number of combinations could be reduced if only two-class mergers were evaluated in each iteration and the process performed repeatedly:

$$\sum_{k=n}^{3}\binom{k}{2} = \binom{n}{2} + \binom{n-1}{2} + \binom{n-2}{2} + \cdots + \binom{3}{2}.$$

As an example, the number of possible two-class mergers, evaluated in the first iteration in DataSet1, is 666. If only one merger is performed, the next iteration evaluates 630 mergers. However, if the end-user decides to perform four mergers, this results in 33 classes in the next merging-iteration, with 528 mergers to evaluate. In spite of this

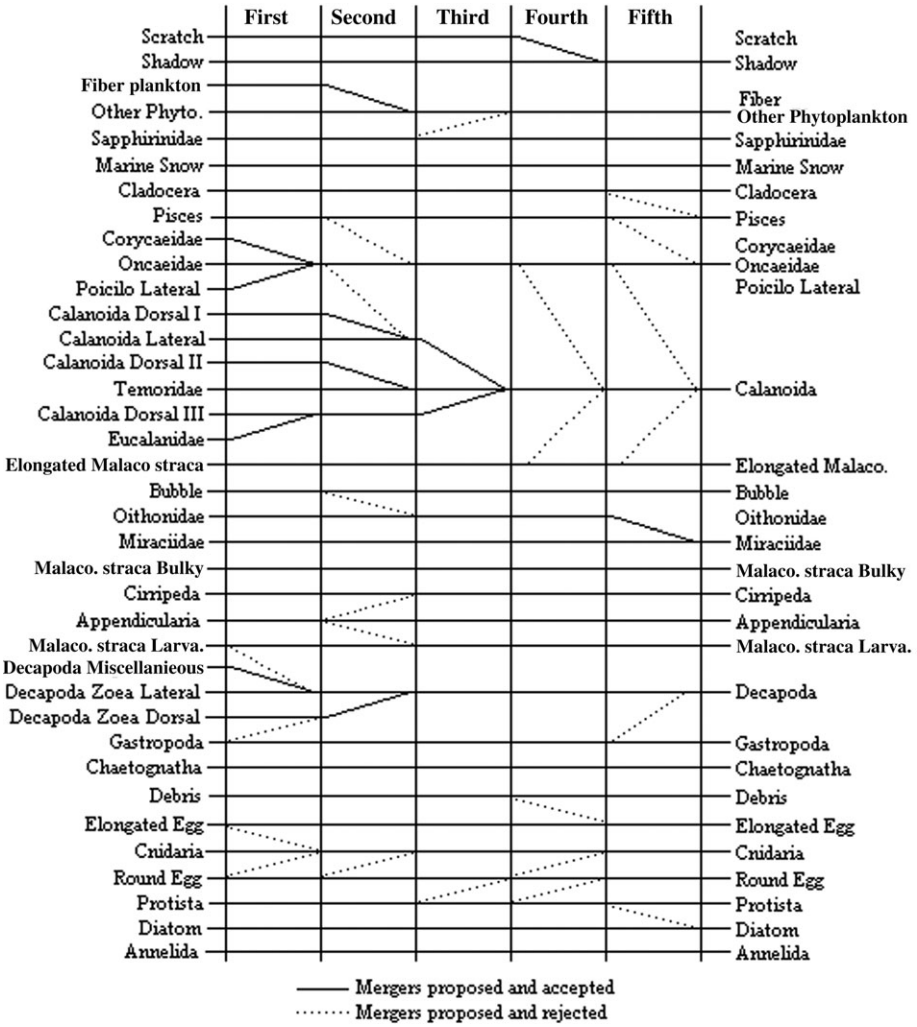

Fig. 2. Graphical representation of accepted mergers by the end-user in straight lines for each iteration. In dotted lines, some machine proposed mergers rejected by the end-user.

reduction in number of evaluations, it remains a computationally expensive task (several hours for DataSet1 first iteration, Table V) that can be reduced using the CM to find a good set of merger candidates instead of trying all two-class mergers (<20 min, Table V). Occasionally, more than one merger per iteration could lead to a lower accuracy. However, this has never been observed in our experiments and several mergers per iteration are selected by the user to speed up the process.

The proposed method application presents a number of benefits: (i) the end-user has a framework within which to accomplish a "trade-off" between the number of classes and performance; (ii) the absence of monotonicity between the number of classes and accuracy can result in improved performance for more detailed data sets. The suggestion of commencing with the most detailed data set benefits from this lack of a strict dependency; (iii) the user can avoid testing mergers that actually decrease performance.

The particular objectives of each end-user's study have an impact on the decision of accepting or rejecting mergers. However, the end-user faces the question of whether the accuracy gains obtained after merging classes are relevant or not. The proposed metrics (accuracy, PRE, TP, FP and the *P*-value) should help in taking such decisions and to evaluate classifiers' effectiveness. The following example using DataSet1 illustrates a possible use of these metrics: the accuracy gain is not significantly higher after the third iteration, so the end-user could make use of the classifier obtained at that step. However, the TP rate of Oithonidae and Miraciidae improves with the classifier obtained after the fifth iteration (Table VII). If these classes were important for the end-user's research, the decision

would be to select the classifier obtained after the fifth iteration. Most of merged classes in all data sets present significant improvements in TP and FP with little variations in the rest of the classes.

The aim of the proposed method is to reduce the end-user's uncertainty, by providing guidance to balance the number of classes and the classification performance. The end-user can initially separate all the identifiable groups, check the decision in terms of automatic classification and then evaluate the proposed changes according to performance (accuracy, PRE, TP, FP and significance of the improvements) and the research objectives. Lastly, the method is independent of any specific machine-learning technique, but sample techniques are selected and a code implementation is provided. Future work will focus on the automation of mergers exploration and on the unbalanced nature of zooplankton data sets.

SUPPLEMENTARY DATA

Supplementary data can be found online at http://plankt.oxfordjournals.org.

FUNDING

J.A.F. research was supported by a Doctoral Fellowship from the Fundación Centros Tecnológicos Iñaki Goenaga. This work has been supported partially by the Etortek, Saiotek and Research Groups 2007–2012 (IT-242-07) programs (Basque Government), TIN2008-06815-C02-01 and Consolider Ingenio 2010 - CSD2007-00018 projects (Spanish Ministry of Science and Innovation) and COMBIOMED network in computational biomedicine (Carlos III Health Institute). This research was partially funded by the project PERFIL (CTM2006-12344-C02-02) from the Spanish Ministry of Education and by the Department of Agriculture, Fisheries and Food of the Basque Country Government. This is contribution number 427, of the Marine Research Division of AZTI-Tecnalia.

ACKNOWLEDGEMENTS

We wish to thank Philippe Grosjean for his help and training, in the use and internal working knowledge of ZooImage. We are grateful to all the anonymous reviewers and the Editor, for their valuable comments that resulted in improvement of the manuscript. Finally, Professor Michael Collins (SOES, University of Southampton, UK and AZTI-Tecnalia, Spain) is acknowledged for his comments on the manuscript.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1965) *Handbook of Mathematical Functions*. Dover, New York.

Alcaraz, M., Saiz, E., Calbet, A. *et al.* (2003) Estimating zooplankton biomass through image analysis. *Mar. Biol.*, **143**, 307–315.

Alpaydin, E. (2004) *Introduction to Machine-Learning.* MIT Press, Cambridge, MA, pp. 17–38, 327–350.

Benfield, M. C., Grosjean, P., Culverhouse, P. *et al.* (2007) RAPID: research on automated plankton identification. *Oceanography*, **20**, 12–26.

Bouckaert, R. R. and Frank, E. (2004) Evaluating the replicability of significance tests for comparing learning algorithms. *Proceedings of the 8th Pacific-Asian Conference on Knowledge Discovery and Data Mining, Australia*. pp. 3–12.

Buckheit, J. B. and Donoho, D. L. (1995) Wavelab and reproducible research. *Technical Report 474*. Department of Statistics, Stanford University.

Culverhouse, P., Williams, R., Reguera, B. *et al.* (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. *Mar. Ecol. Prog Ser.*, **247**, 17–25.

Culverhouse, P., Williams, R., Benfield, M. *et al.* (2006) Automatic image analysis of plankton: future perspectives. *Mar. Ecol. Prog Ser.*, **312**, 297–309.

Donamukkala, R., Huber, D., Kapuria, A. *et al.* (2005) Automatic class selection and prototyping for 3-D object classification. *Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling, IEEE Computer Society.*

Friedman, N., Geiger, D. and Goldszmidt, M. (1997) Bayesian network classifiers. *Mach. Learn.*, **29**, 131–163.

Geisser, S. (1975) The predictive sample reuse method with applications. *J. Am. Stat. Assoc.*, **70**, 320–328.

Grosjean, P. h., Picheral, M., Warembourg, C. *et al.* (2004) Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. *ICES J. Mar. Sci.*, **61**, 518–525.

Hagle, T. M. and Glen, E. M., II (1992) Goodness-of–fit measures for Probit and Logit. *Am. J. Polit. Sci.*, **36**, 762–784.

Hu, Q. and Davis, C. (2006) Accurate automatic quantification of taxa-specific plankton abundance using dual classification with correction. *Mar. Ecol. Prog Ser.*, **306**, 51–61.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, **2**, 1137–1143.

Kohavi, R. and John, G. H. (1997) Wrappers for feature subset selection. *Artif. Intell.*, **97**, 273–324.

Luo, T., Kramer, K., Goldgof, D. B. *et al.* (2004) Recognizing plankton images from the shadow image particle profiling evaluation recorder. *IEEE T. Syst. Man. CY. B*, **34**, 1753–1762.

Mackas, D. L. (1984) Spatial autocorrelation of plankton community composition in a continental shelf ecosystem. *Limnol. Oceanogr.*, **29**, 451–471.

Nadeau, C. and Bengio, Y. (2003) Inference for the generalization error. *Mach. Learn.*, **52**, 239–281.

Pazzani, M. (1996) Searching for dependencies in Bayesian classifiers. In Fisher, D. H. and Lenz, H. (eds), *Learning from data: Artificial Intelligence and Statistics V.* Springer, pp. 239–248.

Schaffer, C. (1993) Selecting a classification method by cross-validation. *Mach. Learn.*, **13**, 135–143.

Steele, J. H. (1989) The ocean 'landscape'. *Landscape Ecol.*, **3**, 185–192.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc.*, **36**, 111–133.

Strathmann, R. R. (1967) Estimating the organic carbon content of phytoplankton from cell volume or plasma volume. *Limnol. Oceanogr.*, **12**, 411–418.

Wiebe, P., Boyd, S. and Cox, J. L. (1975) Relationships between zooplankton displacement volume, wet weight, dry weight, and carbon. *Fish. Bull.*, **73**, 777–786.

Witten, I. H. and Frank, E. (2005) *Data Mining—Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.